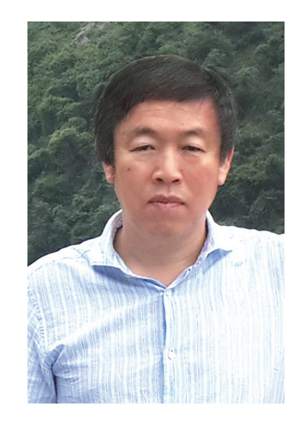
土壤环境大数据：构建与应用



郭书海[1,2,3]　吴　波[1,2,3]　张玲妍[1,2,3]　罗　明[4]

1 中国科学院沈阳应用生态研究所　沈阳　110016
2 污染土壤生物–物化协同修复技术国家地方联合工程实验室　沈阳　110016
3 辽宁省土壤环境大数据工程技术研究中心　沈阳　110016
4 国土资源部土地整治重点实验室　北京　100035

摘要　文章从大数据特征分析入手，分析了国内外大数据在环境领域的发展状况，阐明了我国土壤环境大数据发展的数据基础与瓶颈问题，提出了土壤环境大数据系统的构建方法与技术流程；并根据国家大数据发展战略与土壤环境领域的行业需求，建议统筹建立土壤环境大数据云平台、管理平台和专题应用平台，提供面向区域尺度土壤环境管理、多主体跨介质协同治理和农产品安全保障的公共服务与创新应用产品。

关键词　土壤环境，大数据，数字化管理，多介质协同治理，农产品质量安全

DOI　10.16418/j.issn.1000-3045.2017.02.011

1 大数据

1.1 大数据特点

大数据就是巨量的数据集合，是一种规模大到在获取、存储、管理、分析方面大大超出了传统数据库软件工具能力范围的数据集合，具有海量数据规模、快速数据流转、多样数据类型等特征，需要更强决策力、洞察发现力和流程优化能力的新处理模式才能适应的信息资产[1,2]。

大数据由于数据规模巨大，相比传统数据，有两个明显的特征：

（1）数据属性多样，包括结构化、半结构化和非结构化数据[3,4]。大数据不仅包括数字，还包括文本、图片、音频、视频等多种格式，涵括内容十分丰富，可挖掘属性强，更具潜在应用价值。

（2）数据交互频繁，大规模的数据分析与实时数据挖掘并行[5]。在数据分析中，对于结构化数据，可以遵循一定现有规律[3]，而大数据中半结构化和非结构化数据的分析所遵循

*资助项目：国土资源部公益性行业科研专项（201511082），中科院“一三五”重点突破项目（Y2YZX181YD）
修改稿收到日期：2016年12月26日

的规律是未知的，只能通过综合模拟-假设应答的方式，计算各种可能性的可信度[3]。

大数据的采集主要有三种形式：（1）采集公众信息，进行个性化分析；（2）采集传感器数据，进行专业性预测分析；（3）采集整理综合数据，进行相关性对比分析。

大数据技术领域主要包含数据管理、计算处理和数据分析，其中数据分析是大数据的核心。数据分析经过了若干历史阶段：第一阶段是朴素的数据分析，如占卜、农耕推算等；第二阶段是基于数学科学的数据分析，即采用概率论与统计学等相关数学科学方法，应用计算机技术进行分析；第三阶段是信息技术革命后，对数据进行结构化、数字化处理，开展了基于计算机和数学等技术的集成性分析；第四阶段，即目前的大数据分析，融合了互联网、自动化、计算机、数学科学等技术的融合性数据分析。由此可见，大数据技术中的数据分析是广义概念，不仅包括狭义的数据分析，而且包括巨量数据的深度挖掘。

1.2 大数据应用

大数据目前已经渗透到现代社会的方方面面（表1）。在商业销售领域，各国电商通过公众信息采集，掌握客户网络消费行为与消费特征，进行商品定制生产与精准营销[6]；在智能生产领域，欧美国家已经将实时监测网络做到了终端，利用监测跟踪系统的高频数据，通过积累大量的先验数据，预测用户决策和市场需求，适时调整生产计划；在智慧管理领域，以云 3D GIS 三维地理空间信息引擎及云数据中心为支撑，将各种数据、图表进行分类收集、整理，再经过汇总、分析，并通过发布、反馈、修正等环节，开展跨平台、跨网络、跨终端管理，实现从传统模式向现代管理方式的转变。总之，大数据是信息产业发展到一定阶段的产物，主要来源于公众参与后的投影数据、传感器采集的在线数据和收集整理的多元化综合性数据。

2 环境大数据发展现状

环境领域的大数据目前也处于蓬勃发展阶段，并且显示了广阔的应用前景。

2.1 欧美国家环境大数据发展迅速

由于欧美等国信息化程度较高，大数据基础较好，因此环境领域的大数据发展较为迅速。尤其是美国国家环保局（EPA），已经将环境大数据服务应用于监测网络、数据共享及公共服务[7,8]。在监测网络建设方面，EPA 对企业、污水处理厂、民用设施、采矿作业等享有排污权的设施进行登记，通过唯一“设施标识码”构建排污设施登记数据库，实现跨业务系统和跨库检索[8]。在数据共享方面，EPA 通过环境信息交换中心（Central Data Exchange），实现环境数据快速、有效、安全且精确的实时交换，以此连接美国联邦政府、地方政府、企业及 EPA 各分支单位[9]。在公共服务方面，EPA 通过环保状况数据库（Envirofacts）[8]，以地图可视化的模式，将空气、水、废、毒、辐射、土壤等环保数据系统开放给社会大众，可检索废气排放量、排水许可证、危废处理过程、有毒化学品排放、超基金状态等公众关注信息。

2.2 我国大气环境管理率先采用大数据技术

大气环境数据易于采集和分析，我国的雾霾治理需求又极为迫切，这两个因素的叠加促进了我国在大气环境大数据领域的发展。北京市环保局与IBM公司合作，基于认知计算、大数据分析以及物联网技术的优势，分析空气监测站和气象卫星传送的实时数据流，凭借自学

表 1 大数据采集方式及应用案例

数据采集方式	代表性应用领域	方法与目的	案例
公众信息采集	商业销售	趋势判断与定制服务	Amazon、淘宝、京东等电子商务销售平台
高频传感器采集	生产计划	先验分析与定量生产	西门子智能用电控制、波音飞机检修与制造
多元信息集成	智慧管理	综合分析与优化调度	Inrix 城市交通管理、Google 的 NEST 智能家居

习能力和超级计算处理能力，研发空气质量预测和建模系统，提供未来 72 小时的高精度空气质量预报，实现对北京地区的污染物来源和分布状况的实时监测，即“绿色地平线”项目[9]。“绿色地平线”利用大数据和人工智能，可预测长达 10 天的空气污染状况。城市管理者可以就此采取非常有针对性的措施，比如可以提前改变某些城市的交通模式、控制工业大气污染物的排放等。有了准确的预测，下一步还可以通过 APP（应用）采集很多非结构化的数据，比如天气的规律、科学杂志的内容或者政府报告等等，发展为认知型技术。

由于环境介质、污染物特征、监测手段与历史积累等差异，大数据在环境领域的应用与前景也存在差别（表 2），大气、水、土壤环境大数据的发展特点各异，应针对性的开展大数据构建系统与应用研究。

3 土壤环境大数据发展现状

3.1 土壤环境大数据特点

由于环境研究对象的属性各异，我们能够获得的数据类型也有很大差别。大气环境数据较容易通过传感器进行高频率采集，公众对大气环境质量也有直接和敏感的感官认知，公众参与度高、反馈及时是目前大气环境大数据在环境领域先行一步的客观原因。相比而言，土壤环境质量的变化慢、波动小，污染具有累积性和滞后性的特点，公众没有直接的感官判断能力，也难以进行自动在线监测，人工采样监测的成本更高，因此，在预报预警方面难度较大。但土壤环境质量的变化特点也为大数据发展提供了另一个优势，即针对土壤环境的“源-汇”特性，探索土壤环境质量与各种影响因子的因果关系，通过多元化数据，如整合区域内污染源空间分布数据、污染物排放类别与总量数据、污染扩散的多维途径、环境的消纳能力与空间差异，以及与环境质量相关的背景值图集、各种遥影像资料等，建立基于时空的多维大数据模型。

3.2 土壤环境大数据发展基础

从 20 世纪 80 年代开始，我国开展了多次全国尺度的土壤环境调查，包括全国背景值调查[10]、土壤污染状况调查[11]、多目标地球化学调查[12]、农产品产地环境调查[13]等，此外还形成了超过两百万篇的科研论文与报告。已经积累了以农用地、污染场地和饮用水水源地土壤为重点，涉及局部地区农产品、人群健康等信息的土壤环境基础数据库及衍生数据库（表 3），从数据量上来看，已

表 2　环境领域大数据发展特点

环境领域	介质流动性	污染变化	污染预测模型	静态历史数据	动态采集数据		直接数据量	扩展数据量	数据源扩展方向
					传感器	频率			
大气环境	强	快	成熟	较多	精度高	高	大	较大	人群与环境受体
水环境	强	较快	相对成熟	多	精度高	中	中	少	地表水、地下水
土壤环境	无	慢	基本空白	少	无	无	小	较大	多介质、污染源

表 3　土壤环境大数据的数据基础

数据源	数据产生时间	数据源	数据类型	数据量(TB)
全国土壤环境背景值调查	1975—1980年	调查报告	数值型	1.82×10^{-6}
全国生态环境地球化学调查	1991—1996年	调查图集	数值型	8.00×10^{-6}
全国土壤污染现状调查	2006—2011年	调查报告	数值型	3.27×10^{-5}
全国多目标地球化学调查	2006年至今	报告图集	数值型	2.36×10^{-4}
农产品产地土壤重金属污染普查	2012年至今	调查资料	数值型	4.58×10^{-5}
全国重点污染源调查	1989年至今	结果公告	数值型	1.28×10^{-3}
全国尺度各类调查影像	1975年至今	调查资料	非数值型	4.01×10^{-1}
遥感影像（SPOT、MODIS）	2002年至今	公开数据	非数值型	9.64×10^{2}
各类研究报告与科技论文	1980年至今	公开文献	混合型	5.45×10^{-4}

经基本达到大数据要求，但仍需进行有效数据提取与深度发掘。2016年国务院印发的《土壤污染防治行动计划》[14]将土壤污染调查与监测作为重点，建立每10年开展一次的土壤环境质量状况定期调查制度，建设土壤环境质量监测网络，2020年底前实现土壤环境质量监测点位所有县、市、区全覆盖。这为土壤环境大数据提供了覆盖全国的基础性数据源，为构建样本量巨大性、数据多源性、指标动态性的土壤环境大数据奠定了基础。在此基础上，利用“互联网+”信息互换模式，开展土壤环境数据的摄取与补充，通过数据自我比对、自我更新和自我完善，构建具有我国特色的土壤环境大数据系统，实现土壤环境数字化，以“靶向”服务为目标，为区域性、全国性等不同尺度的土壤环境保护与风险管控提供决策方案。

3.3 土壤环境大数据发展瓶颈

土壤环境大数据发展也存在诸多问题。（1）土壤环境质量监测成本高、周期长，积累数据尚不充分；（2）我国环境监测体系还处于构建阶段，数据种类比较单一，数据分析手段仍处于初级阶段，缺乏数据融合及深度挖掘的方法，亟需构建数据间相关性分析的数学模型；（3）土壤环境质量的管理须基于地理信息系统（GIS），但GIS工具与关系数据库管理系统的扩展能力有限，受限于数据存储模式等诸多瓶颈，导致地理信息系统空间数据自动综合能力与效率低下；（4）GIS的客户机服务器架构决定了数据共享、数据存储、同步性更新及更新效率等能力较弱。因此，应通过技术集成，建立数据驱动的多行业、多学科交叉融合，互利共赢，形成智慧型土壤环境管理数据支撑体系。

4 土壤环境大数据系统构建

大数据具有海量、多样、快速变化的特性，同时海量数据存在价值密度低的特点，这就要求在针对具体问题进行数据分析与价值挖掘时要进行数据的聚合、抽取等预处理工作，以降低计算成本。大数据分析项目经验表明，高可用、可扩展的数据存储架构和灵活、高效的数据分析架构是建设一个完善的大数据分析系统的基本问题。由于土壤环境大数据的采集途径多样，数据来源广泛，因此需首先进行数据融合（Data Blending），再进行集成分析。

数据融合是以智能决策为目标，将多源中的相关数据提取、融合、梳理整合成一个分析数据集（Analytic Dataset）[15]。这个分析数据集是个独立的和灵活的实体，可随数据源的变化重组、调整和更新。

数据融合过程中的多源数据[15]来自于三个方面：（1）基本数据（Primary Data），主要指项目组织者直接采集掌控的内部数据；（2）二级数据（Secondary Data），主要指第三者采集、整理和提供的外部数据；（3）科学数据（Scientific Data），主要指通过科学研究、公式计算和模型估算等获得的数据。这三类数据为系统的建立提供了不同数据信息。在大数据分析项目中，数据科学家需要针对具体问题收集、整理、融合相关的三类数据。

大数据的数据融合与系统构建有5个基本步骤：（1）从多个异构数据源中抽取数据；（2）对数据进行整理和分类；（3）对数据进行清洗；（4）对多元数据进行组合，转换数据并建立数据集；（5）面向具体问题建立数据分析模型。

根据土壤环境大数据的特点，以土壤环境质量为核心的大数据系统，其建立应该遵循以下技术路线（图1）。

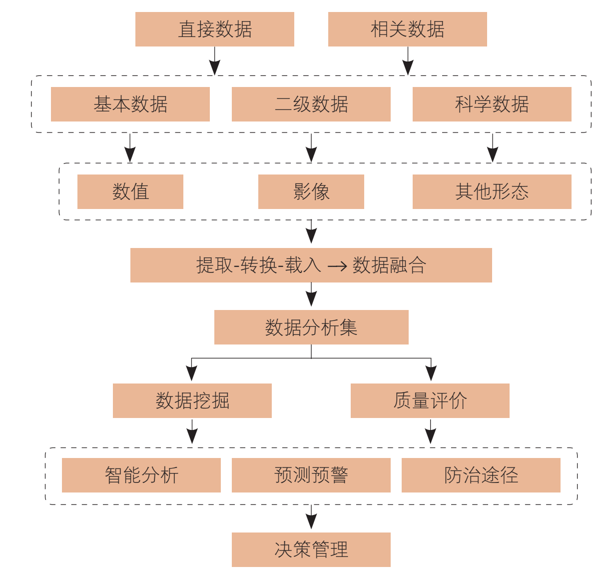


图1 土壤环境质量大数据构建与应用

其中，直接数据指直接表征土壤环境质量的数据，污染物类型、总量、有效态含量等，相关数据指影响土壤环境质量的数据，如土壤理化性质，污染源的空间分布和排放特征、污染物扩散途径、土壤环境的自净能力、水气等相关介质的环境质量特征等，还包括气象资料、水文地质资料、环境影像资料、遥感资料等其他格式的表征类数据。

5 土壤环境大数据发展方向与应用

通过数字土壤环境的大数据集合，搭建保护与防治等专题平台，提供基于土壤环境大数据的公共服务；利用大数据的深度挖掘与知识发现，实现土壤环境的量化管理和多主体跨介质协同治理；面向污染土壤的靶向修复与安全利用，建立保障农产品质量安全的数字化溯源网络，从而保障区域农产品质量安全（图2）。

5.1 提供基于土壤环境大数据的数字化公共服务

国务院印发的《促进大数据发展行动纲要》[16]，要求发展大数据在工业、新兴产业、农业农村等行业领域应用，形成大数据产品体系，完善大数据产业链。《土壤污染防治行动计划》[14]，也要求要利用环境保护、国土资源、农业等部门相关数据，建立土壤环境基础数据库，构建全国土壤环境信息化管理平台，借助移动互联网、物联网等技术，拓宽数据获取渠道，实现数据动态更新。据此，应开展多源数据融合与数字化表征，探索土壤环境质量数据库与多元评估方法、土壤环境质量区域分析与目标控制模型、污染土壤修复的情景分析与决策技术，建立全景式的土壤环境质量分析模式，并在此基础上，根据土壤环境大数据系统需要，统筹建立土壤环境大数据的云平台和专题应用平台，为社会提供基于土壤环境数据的各种数字化公共服务。

5.2 开展面向区域尺度的多主体跨介质协同治理

跨介质环境污染研究是目前国际上最活跃的前沿领域，掌握多介质环境污染的来源、成因、影响和控制尤为重要。单纯进行土壤的污染预防、风险管控和治理修复，已经难以满足社会需求，亟需加强污染源、污染途径和环境承载力等多元化数据的关联分析，进行综合研判，形成跨部委、跨行业的国家或跨区域管理平台，因此，应在传统环境管理的基础上，融合经济社会、基础地理、气象和水文等数据资源，建设基于空间地理信息系统的土壤环境大数据系统，服务于区域性跨介质协同治理。如对于城市“棕色地块”（处于被废弃状态的土地），建立集成基础数据与信息的云服务平台，用于疑似污染地块的历史调查、产业分析、多介质相互影响及环境对策等；对于大尺度土壤环境管理，整合水土气环境监测、矿产资源调查、环境容量分析、区域社会发展状况和产业结构等信息，开展分区、分类、分级保护和治理[17]。

5.3 建立保障农产品质量安全的数字化溯源网络

农产品产地土壤环境质量直接影响农产品安全。我国中南和西南等高背景值区、有色金属矿区、北方大型污灌区，以及长三角、珠三角和京津冀城郊区，土壤污染均较重，严重威胁粮食和蔬菜质量安全。因此，建立精准至地块的农产品产地管理平台，通过编码系统，开展风险预警，为高品质农产品的增值销售和普通农产品安全风险管控提供服务，是未来农用地土壤环境管理的必然趋势。目前实行的农产品溯源方法，只能进行事后处理，将逐渐被事前干预模式所取代或融合。由此可见，在源头上根据现有土壤环境与农产品质量的调查数据，进行深度挖掘，研发以农产品重金属超标风险协同管控为核心的预报预测及决策技术，将成为今后十年内的主要基础性工作。

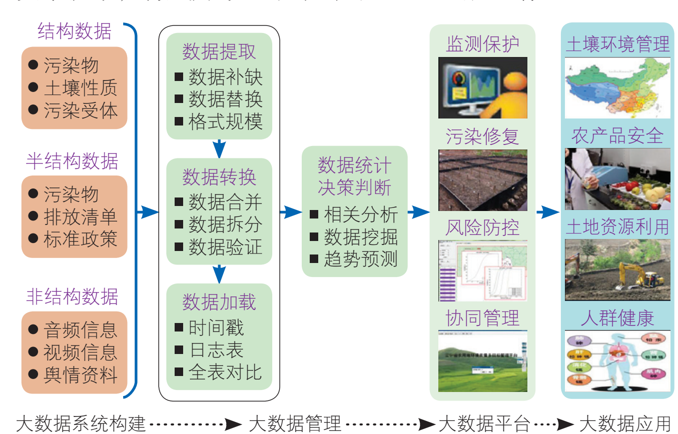


图2　土壤环境大数据体系

6 建议

（1）推进土壤环境数据资源全面整合共享，统筹信息化项目建设管理，建立无偿的基础性国管信息库和有偿的商业性企业信息库，破除数据孤岛。

（2）建立形成环境信息资源中心，实现数据互联互通，形成向平台直接获取为主、部门间数据交换获取为辅的数据共享机制。

（3）研发土壤环境大数据的分析技术，提供公共服务和商业化产品，为区域尺度土壤环境管理、多主体跨介质协同治理和农产品安全保障提供数据与决策支撑。

参考文献

1 Manyika J, Chui M, Brown B, et al. Big data: The next frontier for innovation, competition, and productivity. [2012-10-02]. http://www.mckinsey.com/Ingishts /MGI/Reaearch/Technology-and - Innovation/Big data - The next frontier for innovation

2 维克托·舍恩伯格, 肯尼思·库克耶. 大数据时代. 盛杨燕, 周涛, 译. 杭州: 浙江人民出版社, 2013.

3 Karger D R, Bakshi K, Huynh D, et al. Haystack: A customizable general-purpose information management tool for end users of semistructured data//Proc. Of the CIDR Conf, 2005.

4 IBM. What is big data?. [2012-10-02]. http://www-01.ibm.com/software/data/bigdata/

5 Barwick H. The “four Vs” of Big Data. Implementing Information Infrastructure Symposium. [2012-10-02]. http://www.computerword.com.au/article/396198/iiis-four-vs-big-data/

6 徐国虎, 孙凌. 基于大数据技术的线上线下电商用户数据挖掘流程分析. 中国集体经济, 2012, (30):187-188.

7 USEPA. Facility Registry Service (FRS). [2016-11-29]. https://www.epa.gov/enviro/facility-registry-service-frs

8 USEPA. Envirofacts. [2016-10-21]. https://www3.epa.gov/enviro/

9 环球网科技. IBM推“绿色地平线”计划 助力中国“雾霾战”. [2014-7-8]. http://tech.huanqiu.com/it/2014-07/5051822.html

10 魏复盛, 杨国治, 蒋德珍. 中国土壤元素背景值基本统计量及其特征. 中国环境监测, 1991, 7(1): 1-6.

11 中华人民共和国环境保护部, 中华人民共和国国土资源部. 全国土壤污染状况调查公报. [2014-4-17]. http://www.gov.cn/foot/2014-04/17/content_2661768.htm

12 奚小环. 多目标区域地球化学调查// “十五”重要地质科技成果暨重大找矿成果交流会. 2006.

13 新浪财经.农业部用5年调查农产品产地土壤状况 这些地方污染严重. [2016-11-25]. http://finance.sina.com.cn/roll/2016-11-25/doc-ifxyawxa2737292.shtml

14 中华人民共和国国务院. 土壤污染防治行动计划. [国发〔2016〕31号]. [2016-5-31]. http://www.gov.cn/zhengce/content/2016-05/31/content_5078377.htm

15 韩崇昭, 朱洪艳, 段战胜. 多源信息融合. 北京: 清华大学出版社, 2006.

16 中华人民共和国国务院. 促进大数据发展行动纲要. [国发〔2015〕50号]. [2015-9-5]. http://www.gov.cn/zhengce/content/2015-09/05/content_10137.htm

17 郭书海, 吴波, 李宝林, 等.中国土壤环境区划——原理、方法与实践. 北京: 科学出版社, 2014.

Soil Environmental Big Data: Construction and Application

Guo Shuhai[1,2,3] Wu Bo[1,2,3] Zhang Lingyan[1,2,3] Luo Ming[4]

（1 Institute of Applied Ecology, Chinese Academy of Sciences, Shenyang 110016, China;

2 National-Local Joint Engineering Laboratory of Contaminated Soil Remediation by Bio-physicochemical Synergistic Process, Shenyang 110016, China;

3 Liaoning Engineering Technology & Research Center of Soil Environmental Big Data, Shenyang 110016, China;

4 Key Laboratory of LandConsolidation and Rehabilitation Ministry of Land and Resources, Beijing 100035, China）

Abstract Based on the analysis of big data's characteristics, development situation of environmental big data was illustrated. The data basis and problems of China's soil environmental big data were analyzed. The construction method and process of soil environmental big data were forwarded. According to the national big data strategy, cloud platform, management platform, and application platform of soil environmental big data were established in order to provide the public service and application products for regional scale soil environmental management, multi-media synergetic remediation, and agricultural product quality security.

Keywords soil environmental, big data, digital management , multi-media synergetic remediation, quality safety of agricultural products

郭书海 中科院沈阳应用生态所研究员，污染土壤生物-物化协同修复技术国家地方联合工程实验室主任，入选中科院特聘研究员计划（核心骨干）。研究方向包括污染土壤风险评估与修复、土壤环境数据融合与挖掘等。近年承担“973”“863”、国家重大科技专项、科技部国际合作、国家行业公益专项等课题11项。获得省部级科技奖励9项。E-mail: shuhaiguo@iae.ac.cn

Guo Shuhai Professor, Principal Research Scientist, Institute of Applied Ecology, Chinese Academy of Sciences, and Director of National-Local Joint Engineering Laboratory of Contaminated Soil Remediation by Bio-physicochemical Synergistic Process. His research mainly covers risk assessment and remediation of contaminated soils, and soil environmental data blending and mining. In recent years, he led 11 projects including “973” project, “863” project, national science and technology major project, international cooperation, and national public welfare project, and gained 9 provincial or ministerial science and technology awards. E-mail: shuhaiguo@iae.ac.cn